iDigBio

Integrated Digitized Biocollections

iDigBio is funded by a grant from the National Science Foundation's Advancing Digitization of Biodiversity Collections Program (Cooperative Agreement EF-1115210). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. All images used with permission or are free from copyright.

Digitization Success Depends on Community Involvement

3 February 2016
Scripps Digitization Meeting
La Jolla, CA

Gil Nelson, PhD
Assistant Professor/Research
iDigBio/Institute for Digital Information and Scientific Communication
Florida State University

iDigBio is funded by a grant from the National Science Foundation's Advancing Digitization of Biodiversity Collections Program (Cooperative Agreement EF-1115210). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. All images used with permission or are free from copyright.

Fifteen Thematic Collections Networks (TCNs), 15 PENs

- InvertNet: An Integrative Platform for Research on Environmental Change, Species Discovery and Identification (*Illinois Natural History Survey, University of Illinois*) http://invertnet.org
- Plants, Herbivores, and Parasitoids: A Model System for the Study of Tri-Trophic Associations (*American Museum of Natural History*) http://tcn.amnh.org
- North American Lichens and Bryophytes: Sensitive Indicators of Environmental Quality and Change (*University of Wisconsin – Madison*) http://symbiota.org/nalichens/index.php http://symbiota.org/bryophytes/index.php (plus 2 PENs)
- Digitizing Fossils to Enable New Syntheses in Biogeography - Creating a PALEONICHES-TCN (*University of Kansas*)
- The Macrofungi Collection Consortium: Unlocking a Biodiversity Resource for Understanding Biotic Interactions, Nutrient Cycling and Human Affairs (*New York Botanical Garden*)
- Mobilizing New England Vascular Plant Specimen Data to Track Environmental Change (*Yale University*)
- Southwest Collections of Anthropods Network (SCAN): A Model for Collections Digitization to Promote Taxonomic and Ecological Research (*Northern Arizona University*) http://hasbrouck.asu.edu/symbiota/portal/index.php
- iDigPaleo: Fossil Insect Collaborative: A Deep-Time Approach to Studying Diversification and Response to Environmental Change
- Developing a Centralized Digital Archive of Vouchered Animal Communication Signals (*Cornell University, Laboratory of Orthithology)*
- The Macroalgal Herbarium Consortium: Accessing 150 Years of Specimen Data to Understand Changes in the Marine/Aquatic Environment
- Collaborative: Documenting the Occurrence through Space & Time of Aquatic Non-indigenous Fish, Mollusks, Algae, & Plants Threatening North America's Great Lakes
- Collaborative Research: The Key to the Cabinets: Building and Sustaining a Research Database for a Global Biodiversity Hotspot
- InvertEBase: reaching back to see the future: species-rich invertebrate faunas document causes and consequences of biodiversity shifts
- The Microfungi Collections Consortium: A Networked Approach to Digitizing Small Fungi with Large Impacts on the Function and Health of Ecosystems (MiCC)
- Documenting Fossil Marine Invertebrate Communities of the Eastern Pacific - Faunal Responses to Environmental Change over the last 66 million years (PCMIF)

Advancing Digitization of Biodiversity Collections (ADBC)

To date: 15 TCNs, ~300 unique institutions, 50 states

Information Dissemination Workshops

In March 2012, the iDigBio Steering Committee established a series of preparation-specific digitization training workshops focused on helping collections managers get started with and/or enhance local digitization programs, all to be held at host institutions.

- DROID (Developing Robust Object->Image->Data, May 2012)
- Herbarium digitization (Valdosta State, September 2012)
- Fluid-preserved collections digitization (U. Kansas, March 2013)
- Dried insect collections digitization (Field Museum, April 2013)
- Collections Digitization (West Virginia, ASB, April 2013)
- Imaging fluid-preserved invertebrates (U. Michigan, September 2013)
- Georeferencing Train-the-Trainers (iDigBio, Gainesville, August 2103)
- Paleontology digitization (Yale Peabody Museum, September 2013)
- Small Herbarium Digitization (Florida State University, December 2013)
- Digitization in the South Pacific (Honolulu, March 2014)
- Paleoimaging (Austin, TX, April 2014)
- Small Herbarium Digitization (Boise, Botany 2014, July 2014)
- Leveraging Digitization Knowledge Across Multiple Domains (Santa Barbara, October 2014)
- CT Scanning and Visualization Short Course (University of Texas, February 2015)
- Vertebrate Digitization (Cornell, May 2015)
- The Contribution of Small Natural History Collections in the 21st Century (SPNHC, May 2015)
- Managing Natural History Collections Data for Global Discoverability (Arizona State, September 2015)
- Digitizing Biological Field Stations (Rocky Mountain Biological Laboratory, September 2015)

Information Dissemination Workshops

In March 2012, the iDigBio Steering Committee established a series of preparation-specific digitization training workshops focused on helping collections managers get started with and/or enhance local digitization programs, all to be held at host institutions.

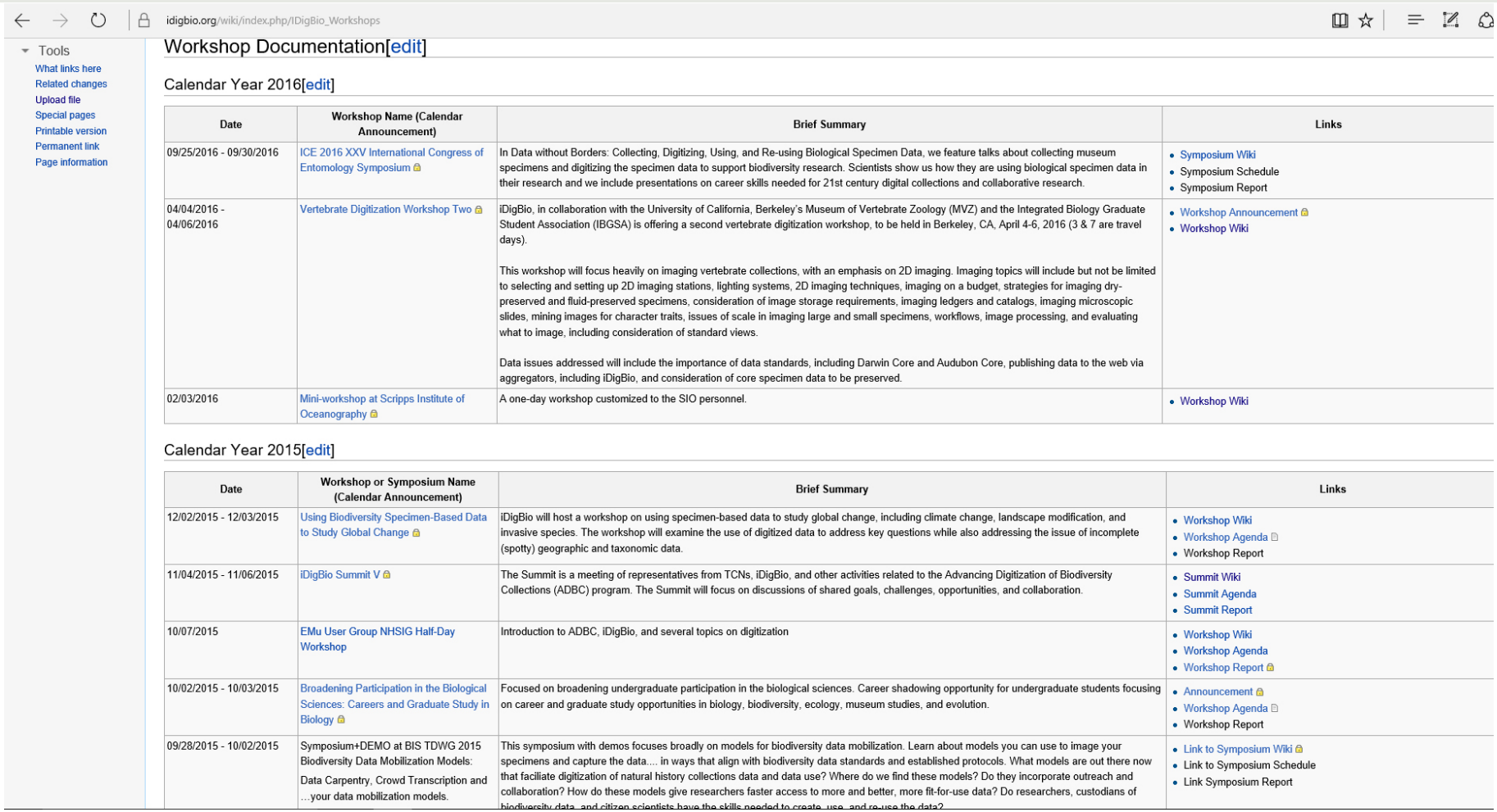
- DROID (Developing Robust Object->Image->Data, May 2012)
- Herbarium digitization (Valdosta State, September 2012)
- Fluid-preserved collections digitization (U. Kansas, March 2013)
- Dried insect collections digitization (Field Museum, April 2013)
- **Collections Digitization (West Virginia, ASB, April 2013)**
- Imaging fluid-preserved invertebrates (U. Michigan, September 2013)
- Georeferencing Train-the-Trainers (iDigBio, Gainesville, August 2103)
- Paleontology digitization (Yale Peabody Museum, September 2013)
- Small Herbarium Digitization (Florida State University, December 2013)
- Digitization in the South Pacific (Honolulu, March 2014)
- Paleoimaging (Austin, TX, April 2014)
- Small Herbarium Digitization (Boise, Botany 2014, July 2014)
- Leveraging Digitization Knowledge Across Multiple Domains (Santa Barbara, October 2014)
- CT Scanning and Visualization Short Course (University of Texas, February 2015)
- Vertebrate Digitization (Cornell, May 2015)
- **The Contribution of Small Natural History Collections in the 21st Century (SPNHC, May 2015)**
- Managing Natural History Collections Data for Global Discoverability (Arizona State, September 2015)
- **Digitizing Biological Field Stations (Rocky Mountain Biological Laboratory, September 2015)**

Information Dissemination Workshops

In March 2012, the iDigBio Steering Committee established a series of preparation-specific digitization training workshops focused on helping collections managers get started with and/or enhance local digitization programs, all to be held at host institutions.

- DROID (Developing Robust Object->Image->Data, May 2012)
- Herbarium digitization (Valdosta State, September 2012)
- Fluid-preserved collections digitization (U. Kansas, March 2013)
- Dried insect collections digitization (Field Museum, April 2013)
- Collections Digitization (West Virginia, ASB, April 2013)
- Imaging fluid-preserved invertebrates (U. Michigan, September 2013)
- Georeferencing Train-the-Trainers (iDigBio, Gainesville, August 2103)
- Paleontology digitization (Yale Peabody Museum, September 2013)
- Small Herbarium Digitization (Florida State University, December 2013)
- Digitization in the South Pacific (Honolulu, March 2014)
- Paleoimaging (Austin, TX, April 2014)
- Small Herbarium Digitization (Boise, Botany 2014, July 2014)
- Leveraging Digitization Knowledge Across Multiple Domains (Santa Barbara, October 2014)
- CT Scanning and Visualization Short Course (University of Texas, February 2015)
- Vertebrate Digitization (Cornell, May 2015)
- The Contribution of Small Natural History Collections in the 21st Century (SPNHC, May 2015)
- **Managing Natural History Collections Data for Global Discoverability (Arizona State, September 2015)**
- Digitizing Biological Field Stations (Rocky Mountain Biological Laboratory, September 2015)

https://www.idigbio.org/wiki/index.php/IDigBio_Workshops



idigbio.org/wiki/index.php/IDigBio_Workshops

Tools
What links here
Related changes
Upload file
Special pages
Printable version
Permanent link
Page information

Workshop Documentation[edit]

Calendar Year 2016[edit]

Date	Workshop Name (Calendar Announcement)	Brief Summary	Links
09/25/2016 - 09/30/2016	ICE 2016 XXV International Congress of Entomology Symposium	In Data without Borders: Collecting, Digitizing, Using, and Re-using Biological Specimen Data, we feature talks about collecting museum specimens and digitizing the specimen data to support biodiversity research. Scientists show us how they are using biological specimen data in their research and we include presentations on career skills needed for 21st century digital collections and collaborative research.	• Symposium Wiki • Symposium Schedule • Symposium Report
04/04/2016 - 04/06/2016	Vertebrate Digitization Workshop Two	iDigBio, in collaboration with the University of California, Berkeley's Museum of Vertebrate Zoology (MVZ) and the Integrated Biology Graduate Student Association (IBGSA) is offering a second vertebrate digitization workshop, to be held in Berkeley, CA, April 4-6, 2016 (3 & 7 are travel days). This workshop will focus heavily on imaging vertebrate collections, with an emphasis on 2D imaging. Imaging topics will include but not be limited to selecting and setting up 2D imaging stations, lighting systems, 2D imaging techniques, imaging on a budget, strategies for imaging dry-preserved and fluid-preserved specimens, consideration of image storage requirements, imaging ledgers and catalogs, imaging microscopic slides, mining images for character traits, issues of scale in imaging large and small specimens, workflows, image processing, and evaluating what to image, including consideration of standard views. Data issues addressed will include the importance of data standards, including Darwin Core and Audubon Core, publishing data to the web via aggregators, including iDigBio, and consideration of core specimen data to be preserved.	• Workshop Announcement • Workshop Wiki
02/03/2016	Mini-workshop at Scripps Institute of Oceanography	A one-day workshop customized to the SIO personnel.	• Workshop Wiki

Calendar Year 2015[edit]

Date	Workshop or Symposium Name (Calendar Announcement)	Brief Summary	Links
12/02/2015 - 12/03/2015	Using Biodiversity Specimen-Based Data to Study Global Change	iDigBio will host a workshop on using specimen-based data to study global change, including climate change, landscape modification, and invasive species. The workshop will examine the use of digitized data to address key questions while also addressing the issue of incomplete (spotty) geographic and taxonomic data.	• Workshop Wiki • Workshop Agenda • Workshop Report
11/04/2015 - 11/06/2015	iDigBio Summit V	The Summit is a meeting of representatives from TCNs, iDigBio, and other activities related to the Advancing Digitization of Biodiversity Collections (ADBC) program. The Summit will focus on discussions of shared goals, challenges, opportunities, and collaboration.	• Summit Wiki • Summit Agenda • Summit Report
10/07/2015	EMu User Group NHSIG Half-Day Workshop	Introduction to ADBC, iDigBio, and several topics on digitization	• Workshop Wiki • Workshop Agenda • Workshop Report
10/02/2015 - 10/03/2015	Broadening Participation in the Biological Sciences: Careers and Graduate Study in Biology	Focused on broadening undergraduate participation in the biological sciences. Career shadowing opportunity for undergraduate students focusing on career and graduate study opportunities in biology, biodiversity, ecology, museum studies, and evolution.	• Announcement • Workshop Agenda • Workshop Report
09/28/2015 - 10/02/2015	Symposium+DEMO at BIS TDWG 2015 Biodiversity Data Mobilization Models: Data Carpentry, Crowd Transcription and ...your data mobilization models.	This symposium with demos focuses broadly on models for biodiversity data mobilization. Learn about models you can use to image your specimens and capture the data.... in ways that align with biodiversity data standards and established protocols. What models are out there now that facilitate digitization of natural history collections data and data use? Where do we find these models? Do they incorporate outreach and collaboration? How do these models give researchers faster access to more and better, more fit-for-use data? Do researchers, custodians of biodiversity data, and citizen scientists have the skills needed to create, use, and re-use the data?	• Link to Symposium Wiki • Link to Symposium Schedule • Link Symposium Report

Financial support
Community planning teams

Have accommodated 2,600+ participants from 500 unique institutions in 75+ workshops. ~20/year.

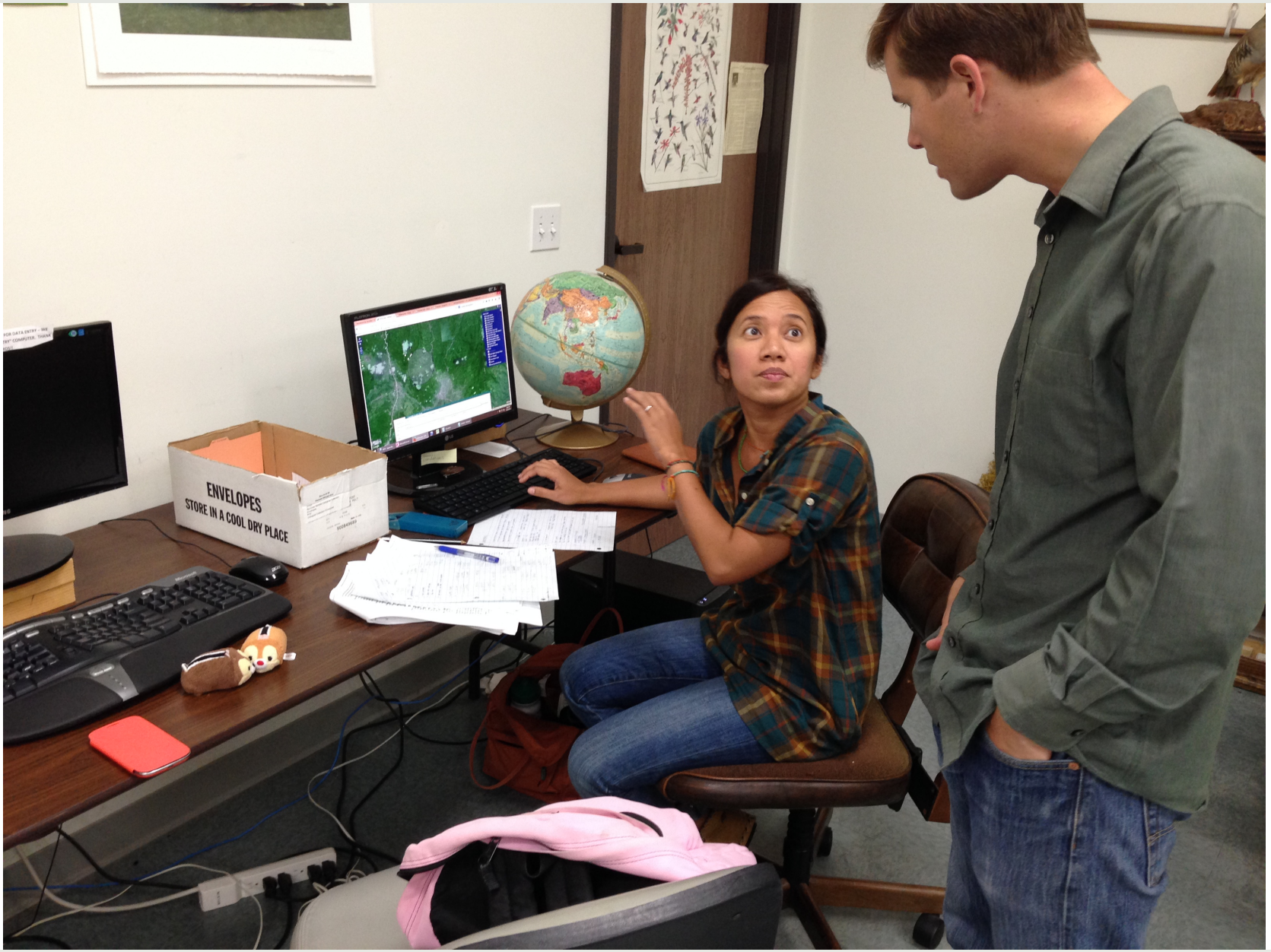
Venues Large and Small

Yale Peabody Museum


GOTHIC
site of
The Rocky Mountain
Biological Laboratory
founded 1928

Demos and Hands on, when possible

Western Foundation of Vertebrate Zoology



Product-oriented Workshops

- Augmenting OCR Hackathon (Ft. Worth, February 2013)
- Original Source Materials Digitization (Yale Peabody Museum, March 2014)
- Recruiting and Retaining Small Collections in Digitization (Mt. Pleasant, MI, April 2014)
- CitScribe Hackathon (iDigBio, Gainesville, December 2013)
- Education and Outreach (iDigBio, Gainesville, January 2014)
- Workflows for Herbarium Digitization (Valdosta State, January 2015)
- Coding Phenological Data from Herbarium Sheets (Berkeley, March 2016)

Wikis
Working groups
Webinars
Listservs

idigbio.org/wiki/index.php/IDigBio_Working_Groups

Forming or Dissolving a Working/Interest Group[edit]

Working Groups and Interest Groups are allowed to form and dissolve organically according to current needs. No formal proposal to iDigBio is required to form or dissolve a Working/Interest Group; simply create and complete the basic information about the group on this page. The Working/Interest Group leader is responsible for maintaining the accuracy of the information on this page.

Active Working Groups[edit]

Augmented Reality Public Education/Outreach Working Group (ARPEO)[edit]

Background: In 2014, the iDigBio Steering Committee voted to fund the production of a set of augmented reality (AR) flashcards similar to those portrayed in https://www.youtube.com/watch?v=STc8Nsx36MI for the ADBC community's public outreach.

Data Field	Details
Status	Active
Members	• Anne Basham, Arizona State University, Project Lead • Austin Mast, iDigBio, Florida State University, Project Co-Lead • Neil Cobb, Northern Arizona University, Southwest Collections of Arthropods Network • Chris Dietrich, University of Illinois at Urbana-Champaign, InvertNet TCN • Nico Franz, Arizona State University, Southwest Collections of Arthropods Network TCN • Andrew Hipp, Morton Arboretum, Great Lakes Invasives TCN • Melissa Islam, Denver Botanical Gardens, Macrofungi Collection Consortium TCN • Talia Karim, University of Colorado Museum of Natural History, Fossil Insect Collaborative TCN • Jennie Kluse, Louisiana State University, Lichens & Bryophytes TCN • Christopher Marshall, Oregon State University • Brendan Morris, University of Illinois at Urbana-Champaign, InvertNet TCN • Katja Seltmann, Cheadle Center, University of California, Santa Barbara • Petra Sierwald, Field Museum of Natural History, InvertEBase TCN • Dena Smith, University of Colorado Museum of Natural History, Fossil Insect Collaborative TCN • Katie Stanley, iDigBio, Florida State University • Gavin Svenson, Cleveland Museum of Natural History, InvertEBase TCN • Patrick Sweeney, Yale Peabody Museum, New England Vascular Plants TCN
Mission & Scope	The working group will produce one card per TCN for a desktop/mobile AR software application. Each card will convey information about the TCN and the value of its targeted biodiversity research specimens and will provide augmented reality buttons that link to further information on the web. For example, a card with a 2-D herbarium specimen image on it could spring to life through the augmented reality viewer, producing a 3-D model of the fresh flower of the species that can be manipulated by turning the card. Combining the herbarium specimen image and an insect card image in the viewer could potentially launch a pollination video—a response that would not occur with either card alone. The goal is to have an exciting interactive tool for formal and informal education initiatives while building collaboration on the education and outreach front within our ADBC community.
Overlap with Other Working Groups	
Links	Wiki

Augmenting OCR (aOCR)[edit]

Background: The Augmenting OCR working group formed from the first iDigBio Summit in 2011 out of a community consensus wish to attempt improving OCR output and algorithm tools to speed up the digitization process. Initial group members were suggested by those present at Summit 2011.

Our current working group met, in person, at a workshop in October 2012 to learn more about each other's use of OCR and OCR output, plan a February 2013 hackathon, develop content for our iSchools 2013 iConference workshop panel, and learn more about the latest developments in handwriting recognition. We put together iDigBio's first hackathon and presented our working group plans at iConference13. In December of 2013, at the iDigBio CITScribe hackathon, the aOCR working group built some working demos of visualization tools that facilitate searching of OCR output. These tools present novel ways to create record sets for citizen scientists, researchers, and data validators too. A recorded demonstration iDigBio webinar: Visualize Text Data From OCR Output on 22 January 2014 presented live demos of these OCR output visualization tools to show how they fit in a digitization workflow.

idigbio.org/wiki/index.php/OBFS_Field_Station_Digitization

Wiki Home
Workshop Summaries
Working Group List
Specimen Portal

iDigBio Data Ingestion
Ingestion Queue Dashboard
Published data
Ingestion Guidance
Data API
Digitization Resources

iDigBio Working Groups

iDigBio Research

Navigation Tools

Tools
What links here
Related changes
Upload file
Special pages
Printable version
Permanent link
Page information

Page Discussion

Read Edit View history

OBFS Field Station Digitization

This wiki supports the joint iDigBio/OBFS Field Station Digitization workshop held at the OBFS annual meeting, 16 September 2015, at Rocky Mountain Biological Laboratory.

Contents [hide]

Digitization Workshop Resources Home[edit]

Field Station Interest Group[edit]

- Link to Field Station Interest Group Wiki

Workshop Documentation[edit]

- Workshop announcement
- Agenda

Workshop Presentations[edit]

- Introduction to iDigBio, Gil Nelson, iDigBio
- Challenges and emerging solutions to building scientific contributions based on digitization of pollinator/host ("flower visitor") collection data, Hilary Swain, Archbold Biological Station
- Using collections data for research and education, Rick Williams, Rocky Mountain Biological Laboratory
- How digitization will help your field station's outreach program, Molly Phillips, iDigBio Education and Outreach Coordinator
- Establishing a field station digitization program: the Sagehen experience
- Engaging citizen scientists through digitized collections, Faerthen Felix, Sagehen Creek Field Station
- Collections funding available through NSF, Roland Roberts, NSF CSBR and ADBC
- Field Station Collection Digitization Workflow, Misha Leong, California Academy of Sciences
  - Field Station Digitization Workflow handout
- Free, open source database options and decisions (including Specify & Symbiota), Gil Nelson, iDigBio
- Guided discussion: Summarizing the grand challenges confronting field stations and how digitization might influence overcoming these despite limited resources, David Jennings, iDigBio

Workflow Resources[edit]

- CalBug Workflows
- iDigBio Workflows at iDigBio
- iDigBio in Word format for download at GitHub

Mobilizing Dark Data
Advancing Scientific Discovery

In an early press release announcing the first round of Advancing the Digitization of Biodiversity Collections (ADBC) awards (July 8, 2011), the National Science Foundation (NSF) several times referenced the importance of what it called "**dark data**"—data that are essentially inaccessible to most biologists, ecologists, policy-makers, the general public, and other scientists.

The longest tail of these "dark data" may well be locked up in small collections that lack sufficient resources to mobilize them for broad use.

Tall Timbers Research Station
Lucien Harris
Butterflies of Georgia
Lepidoptera Collection


iDigBio
Integrated Digitized Biocollections
Home
About SCNet
Webinar Series
Workshops and Symposia
Site map
Small Collections Network
Serving, Supporting, Connecting Small Natural History Collections
Quick Links
Blogs and News
Listserv
Webinar Series
NANSH Webinars and Meetings
Webinar Recordings
Workshops and Symposia
Working Groups
Related Wikis and Links
Relevant Papers & Publications
Introduction to SCNet's Webinar Series
SCNet and iDigBIo are pleased to announce a series of webinars centered on supporting small collections and establishing SCNet as a collaborative resource for small collections and the professionals who manage them. Each webinar in this series will be held 3:00-4:00 p.m. EST on the dates shown below. Meetings are virtual and accessible online at https://idigbio.adobeconnect.com/scnet. No special software outside of an internet browser is required to access the virtual meeting room.
Read more
Webinar Recording - Transcribing Specimens into Symbiota: a practical approach
You can access the webinar recording here:
http://idigbio.adobeconnect.com/p5kbxeuc9k5/
Acces the chatbox entries here.
Read more
Webinar Recording - Achieving the Maximum Potential of Small University Collections: a Model in Digitization, Education, & Outreach
You can access the webinar recording here: http://idigbio.adobeconnect.com/p7xejclok57/
Read more
Follow SCNet on Twitter
Tweets
Follow
Sagehen Creek @SagehenCreekFS 19 Sep
Meeting season! #GNOMO, #iDigBio, and #OBFS at Rocky Mountain Biological Laboratory in Gothic, CO. Erica Krimmel... fb.me/1P7d6rekG
Retweeted by Gil Nelson
Expand
STEPPE @deeptimerocks 21 Sep
Don't miss this webinar by STEPPE partners! iDigPaleo, a great new community resource. fb.me/3Yc8ybdgB
Retweeted by Gil Nelson
Expand
Gil Nelson @iDigGilNelson 10 Sep
Bringing dark data into the light: Best practices for herbarium digitization. eurekalert.org/pub_releases/2... @iDigBio
Show Summary
Rebecca Baldwin @rebbaldwin 27 Aug

Google iDigBio listservs

Current Listservs[edit]

iDigBio Listserv[edit]

Resource: IDIGBIO-L@lists.ufl.edu

Purpose: Community communication and announcements.

To add yourself to the list, email listserv@lists.ufl.edu with the following command in the body of the email:

subscribe idigbio-L first_name last_name

e.g., subscribe idigbio-L Jane Doe

API Users[edit]

Resource: IDIGBIO-API-USERS-L@LISTS.UFL.EDU

Purpose: This list is for announcements about the iDigBio's programming interfaces and ways of programmatically using iDigBio data.

Browse through archives of this list

To add yourself to the list, email listserv@lists.ufl.edu with the following command in the email:

subscribe IDIGBIO-API-USERS-L first_name last_name

e.g., subscribe IDIGBIO-API-USERS-L Jane Doe

Augmenting OCR Working Group[edit]

Resource: IDIGBIO-AOCR-L@LISTS.UFL.EDU

Purpose: This working group is attempting to improve OCR output and algorithm tools to speed up the digitization process.

Browse through archives of this list

To add yourself to the list, email listserv@lists.ufl.edu with the following command in the email:

subscribe IDIGBIO-AOCR-L first_name last_name

e.g., subscribe IDIGBIO-AOCR-L Jane Doe

Biodiversity Informatics Management Working Group[edit]

Resource: IDIGBIOBIM-L@LISTS.UFL.EDU

Purpose: This working group discusses issues related to biodiversity informatics management.

Citizen Science & Crowdsourcing Listserv[edit]

Resource: IDIGBIO-CITIZENSCIENCE-L@LISTS.UFL.EDU

Purpose: Using citizen science resources for digitization and exploring partenerships with crowdcourcing service providers.

Browse through archives of this list

To add yourself to the list, email listserv@lists.ufl.edu with the following command in the email:

subscribe IDIGBIO-CITIZENSCIENCE-L first_name last_name

e.g., subscribe IDIGBIO-CITIZENSCIENCE-L Jane Doe

iDigBio
Integrated Digitized Biocollections

Gaps in Biodiversity Data: Challenges for Digitization

iDigBio is funded by a grant from the National Science Foundation's Advancing Digitization of Biodiversity Collections Program (Cooperative Agreement EF-1115210). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. All images used with permission or are free from copyright.

The Nature of Gaps in the Availability of Digitized Data

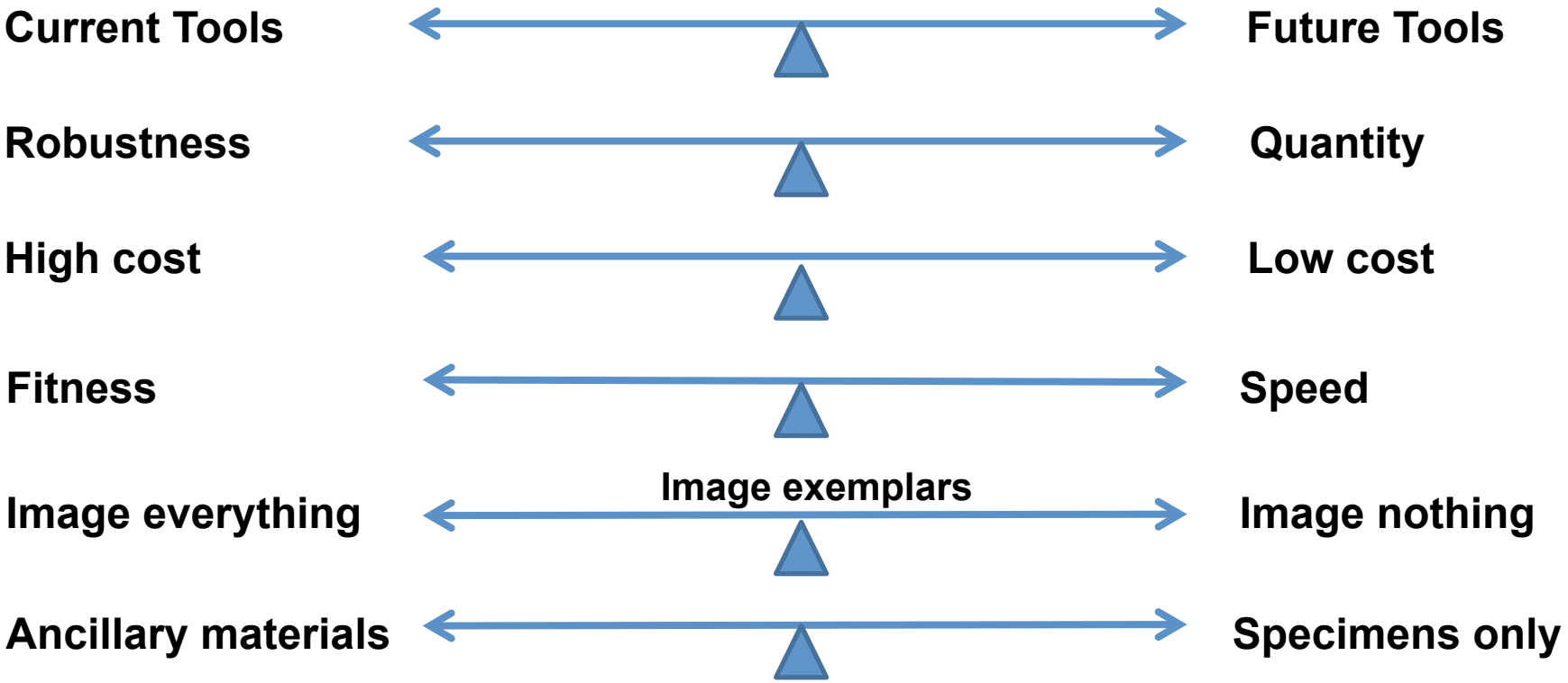
At the institutional level:

- **Strategically ignored data in collections that are otherwise digitized**
  - **Specimen selection**
  - **Data selection**
- **Specimens not yet in the digitization pipeline; differs by institution**

At the community level:

- **Unfunded collections**
- **Small collections**
- **Institutions that choose not to aggregate their data**

Digitization Decision Continua that Influence or Result in Data Gaps



The Nature of Gaps in the Availability of Digitized Data

At the institutional level:

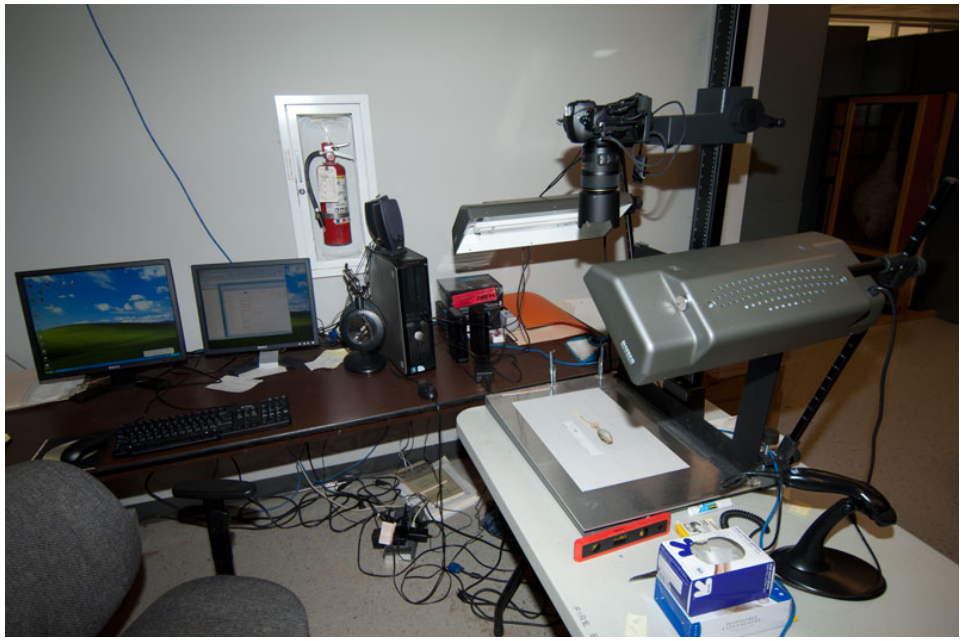
- **Strategically ignored data in collections that are otherwise digitized**
  - **Specimen selection**
  - **Data selection**
- **Specimens not yet in the digitization pipeline; differs by institution**

At the community level:

- **Unfunded collections**
- **Small collections**
- **Institutions that choose not to aggregate their data**

The completeness of digitized biodiversity data that get to any aggregator depends on:

- **what gets digitized,**
- **how it gets digitized,**
- **and what is selected to be shared.**



All dependent on institutional decisions.

Four Basic Assumptions about Data Gaps Directly Attributable to Digitization Practices

The availability of comprehensive, robust, and complete digital datasets from biodiversity specimens is directly dependent on the assumptions made when designing and implementing digitization protocols.

"Missing digital" data for any particular specimen is more likely traced to intentional decisions than to a lack of data to be digitized.

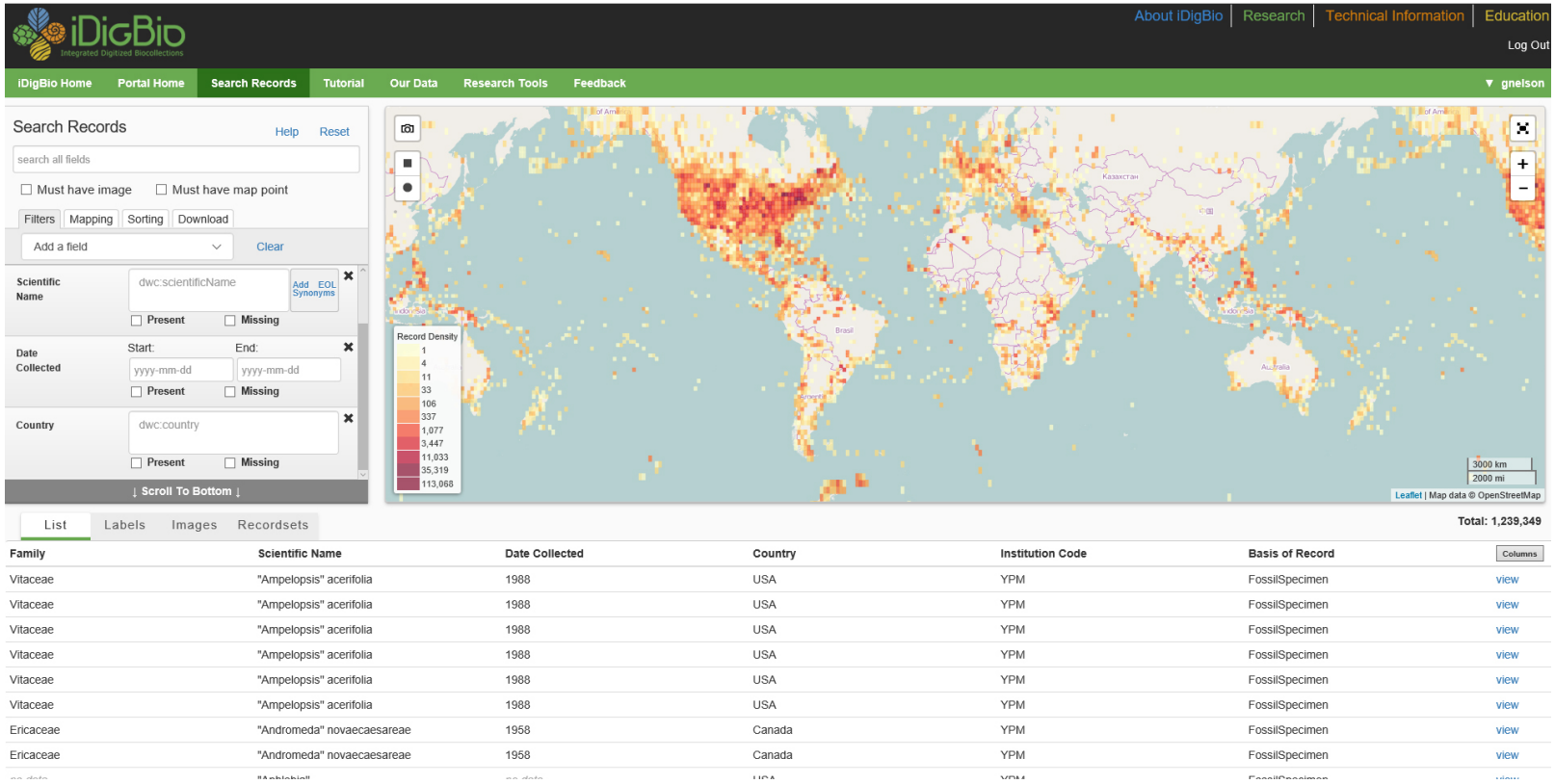
We are still very early in the biodiversity specimen digitization adventure.

Small collections that may contain the darkest of dark data are underrepresented in digitization projects.

Gaps and biases are often designed into digitization protocols for several important reasons:

- **Funder expectations**
- **Tendency toward low cost per specimen**
- **What place imaging holds in the process**

Unevenness in data in the iDigBio portal reflects our mission to accept and ingest all contributed biodiversity specimen data with few restrictions or requirements.



Family	Scientific Name	Date Collected	Country	Institution Code	Basis of Record	
Vitaceae	"Ampelopsis" acerifolia	1988	USA	YPM	FossilSpecimen	view
Vitaceae	"Ampelopsis" acerifolia	1988	USA	YPM	FossilSpecimen	view
Vitaceae	"Ampelopsis" acerifolia	1988	USA	YPM	FossilSpecimen	view
Vitaceae	"Ampelopsis" acerifolia	1988	USA	YPM	FossilSpecimen	view
Vitaceae	"Ampelopsis" acerifolia	1988	USA	YPM	FossilSpecimen	view
Vitaceae	"Ampelopsis" acerifolia	1988	USA	YPM	FossilSpecimen	view
Ericaceae	"Andromeda" novaecaesareae	1958	Canada	YPM	FossilSpecimen	view
Ericaceae	"Andromeda" novaecaesareae	1958	Canada	YPM	FossilSpecimen	view


Records
6000000
5000000
4000000
3000000
2000000
1000000
0
0.095238097
0.269841284
0.285714298
0.301587313
0.317460328
0.333333343
0.349206358
0.365079373
0.380952388
0.396825403
6
17
18
19
20
21
22
23
24
25
Data Quality Score

Gaps Due to Differential Strategies for Data Enrichment

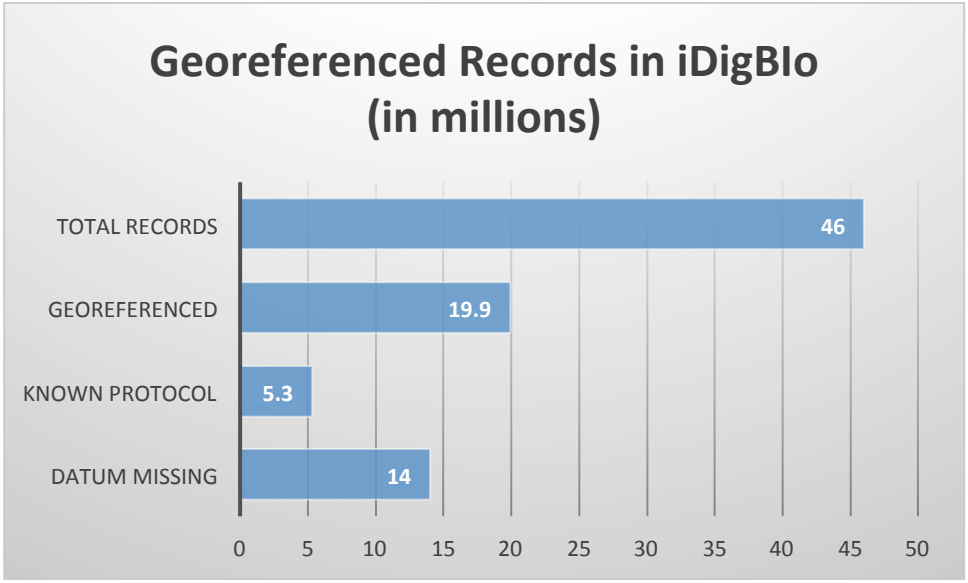
Coding and normalizing:

- Phenological data/anomalies/outliers
- Habitat descriptions/nomenclature/parsing
- Density/abundance/habitat health
- Morphological characteristics and variation

Georeferencing assumptions and inconsistencies:

- Protocol
- Resolution inconsistencies
  - Geographic centroids (county, park, state)
  - Label data
  - Datum
  - Method
- Documentation

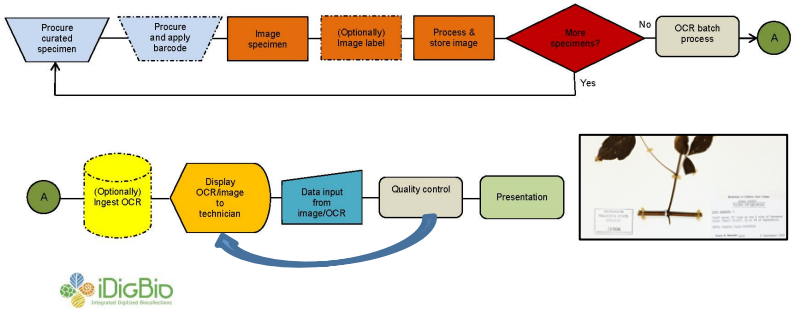
Georeferencing Consistency

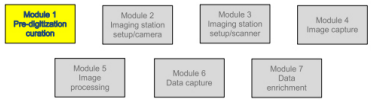


Closing Digitization Gaps

O2I2D(1)—Existing Specimen Workflow Using Optical Character Recognition: Object to Image to Data

This workflow is designed to capture images of existing specimens, pass the images through optical character recognition (OCR) software, and use the combination of image and OCR output to capture data. There are variations on this workflow. For example, depending on preparation type, barcodes are sometimes applied inline as the step immediately previous to imaging (shown optionally below) and other times en masse within an independent step during which several dozen or several hundred barcodes are applied in preparation for imaging. OCR may also occur in various ways: 1) in batch (as shown below), with numerous images being processed following the close of one or more imaging sessions, 2) "on the fly" as a record and its associated image are loaded for data entry, or 3) one image at a time as a step immediately following the imaging of each specimen. OCR output may be ingested into a field in the database (shown optionally below), stored as individual text files within the computer's file system, or virtually processed at the time the image is presented to the data entry technician. The presentation of images and OCR to data entry technicians occurs in a single interface in which database fields, OCR output, and specimen image are simultaneously visible. Pre-digitization curation and annotation is particularly important in this workflow to ensure that the current nomenclature to be used in data entry is obvious and clearly visible in the image and/or OCR output.





Module 1: Pre-digitization Curation Task List

Task ID	Task Description	Explanations and Comments	Resources
T1	Apply storage locator barcodes to storage locations (rooms, cabinets, shelves, folders, drawers, etc).	Most useful when systematically digitizing an entire collection. Otherwise potentially helpful with herbarium inventory. May be less helpful for collections that are digitizing in random order or only portions of the collection related to specific projects, or with significant separation between the pre-digitization curation, databasing, and image capture modules.	Barcodes, QRcode, DataMatrix.
T2	Select specimens to digitize.	For herbaria, this often includes all specimens. Where this is not the case, selection should follow the institution's pre-determined digitization policies or project management plan.	Digitization policy manual or project management plan.
T3	Associate/insert machine readable barcodes/documents with/into folders.	Some institutions create machine readable documents to gather data at the cabinet and/or folder level. Documents might contain such information as family, higher geography, and current identification ("filed-as name"). These data will be read and associated with individual collection records in Module 4, T1 or Module 7.	QRcodes, DataMatrix, 1D barcode, or OCR-readable documents for insertion into specimen folders.

iDigBio has made gains in facilitating the development of digitization workflows in several communities.

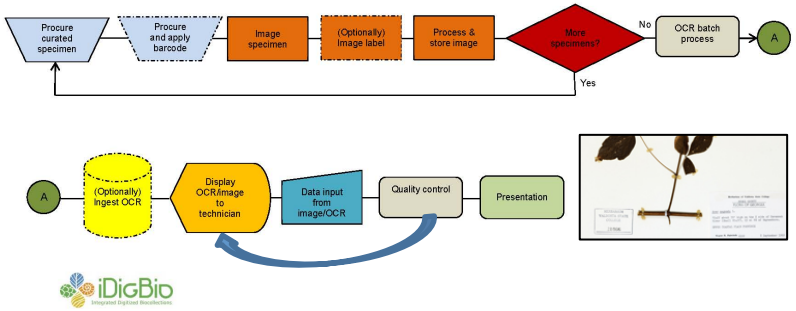
Involve the entire collections community.

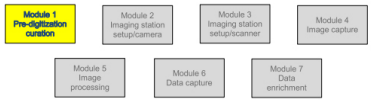
The challenges of institutional variation has encouraged our working groups to provide maximum accommodation via the development of modular, more or less “plug and play” approaches that preserve institutional flexibility.

Closing Digitization Gaps

O2I2D(1)—Existing Specimen Workflow Using Optical Character Recognition: Object to Image to Data

This workflow is designed to capture images of existing specimens, pass the images through optical character recognition (OCR) software, and use the combination of image and OCR output to capture data. There are variations on this workflow. For example, depending on preparation type, barcodes are sometimes applied inline as the step immediately previous to imaging (shown optionally below) and other times en masse within an independent step during which several dozen or several hundred barcodes are applied in preparation for imaging. OCR may also occur in various ways: 1) in batch (as shown below), with numerous images being processed following the close of one or more imaging sessions, 2) "on the fly" as a record and its associated image are loaded for data entry, or 3) one image at a time as a step immediately following the imaging of each specimen. OCR output may be ingested into a field in the database (shown optionally below), stored as individual text files within the computer's file system, or virtually processed at the time the image is presented to the data entry technician. The presentation of images and OCR to data entry technicians occurs in a single interface in which database fields, OCR output, and specimen image are simultaneously visible. Pre-digitization curation and annotation is particularly important in this workflow to ensure that the current nomenclature to be used in data entry is obvious and clearly visible in the image and/or OCR output.





Module 1: Pre-digitization Curation Task List

Task ID	Task Description	Explanations and Comments	Resources
T1	Apply storage locator barcodes to storage locations (rooms, cabinets, shelves, folders, drawers, etc).	Most useful when systematically digitizing an entire collection. Otherwise potentially helpful with herbarium inventory. May be less helpful for collections that are digitizing in random order or only portions of the collection related to specific projects, or with significant separation between the pre-digitization curation, databasing, and image capture modules.	Barcodes, QRcode, DataMatrix.
T2	Select specimens to digitize.	For herbaria, this often includes all specimens. Where this is not the case, selection should follow the institution's pre-determined digitization policies or project management plan.	Digitization policy manual or project management plan.
T3	Associate/insert machine readable barcodes/documents with/into folders.	Some institutions create machine readable documents to gather data at the cabinet and/or folder level. Documents might contain such information as family, higher geography, and current identification ("filed-as name"). These data will be read and associated with individual collection records in Module 4, T1 or Module 7.	QRcodes, DataMatrix, 1D barcode, or OCR-readable documents for insertion into specimen folders.

Adoption of discipline-consensus workflows.

Community agreement on the essential core data requirements that should drive digitization workflows and contribute to research.

Agree on sets of community-based priorities for addressing current data gaps.

iDigBio

Integrated Digitized Biocollections